Evaluating and developing Ingest workflows with OAIS and PAIMAS at the GESIS Data Archive for the Social Sciences

Astrid Recker, Natascha Schumann; GESIS Leibniz Institute for the Social Sciences; Cologne, Germany

Abstract

As part of its current efforts in audit and certification, the GESIS Data Archive for the Social Sciences carried out a mapping to the OAIS reference model to support a systematic analysis of archive workflows and procedures. The mapping helped us to identify areas and processes with potential for improvement. Among these, the Ingest functional entity is of particular importance and complexity. Thus we continued our work by further analyzing our Ingest workflows and processes with the help of the PAIMAS standard, placing particular emphasis on the interfaces employed in the communication with internal and external stakeholders. Discussing these results with involved staff members helped us to identify problematic areas and develop a strategy for the future development of this functional entity at the GESIS Data Archive.

Evaluating Ingest workflows at the GESIS Data Archive

The GESIS Data Archive is part of GESIS – Leibniz-Institute for the Social Sciences, Germany's biggest research-based social sciences infrastructure institution. Founded in 1960, the archive is one of the oldest archives in Germany to actively curate and preserve digital research data for the long term. In an effort to create more transparency, the archive is currently undertaking a series of tiered certification and audit procedures in accordance with the European Framework for Audit and Certification of Digital Repositories [1]. To support the first step in this process, the application for the Data Seal of Approval [2], we carried out a functional mapping between the archive and the OAIS reference model [3]. Driven by questions such as "Are we OAIS compliant?" but also "Where and how do we differ?", the mapping allowed us to gain a systematic overview of the responsibilities and functions that we fulfill as an archive [4]. However, this initial mapping also created further questions for us.

Thus, while OAIS clearly defines interfaces between different functional entities, as a reference model it does not specify how a particular function is to be fulfilled (that is, for example, whether it is to be carried out by a machine or a human being), how it should be triggered, or how – rather than what – one functional entity should communicate to another. The mapping made it clear that in a next step we would have to look at these interfaces in more detail, as it had become apparent that while most of the interfaces between functions and functional entities are in place, how these operate requires further investigation and – potentially – improvement. In particular, the desired degree of automatization, regularization, and standardization needed to be determined.

The OAIS functional entity where such further investigation is particularly relevant for us as a social science data archive is Ingest which "provides the services and functions to accept Submission Information Packages (SIPs) from Producers . . . and prepare the contents for storage and management within the Archive" [5]. The GESIS Data Archive holds many different collections submitted to the archive through different channels – for example, a considerable amount of data is submitted by GESIS's Research Data Centers (RDCs) [6], which also create comprehensive value-added services for the studies they submit. The main focus of these RDCs is on monitoring society and social change in Germany as well as on international comparative survey research and election studies.

As part of Ingest, the archive carries out extensive quality controls and data processing, an aspect that in its extensiveness is not well-accounted for by OAIS. For this reason, we turned to the Producer-Archive Interface Methodology Abstract Standard (PAIMAS) [7] and used it to map and describe our Ingest workflows more fully.

The mapping process was accompanied by discussions with the staff members responsible for Ingest to better understand potential problems or bottlenecks in the workflow and to explore possibilities for the automatization of certain workflows. In particular, the following points showed up as problematic in the discussions:

- Although the internal interfaces between the Research Data Centers and Ingest and the responsibilities are well-defined, the actual Ingest process is complicated by the fact that often large amounts of data are submitted at a time. These can be composed and structured very differently depending on the survey. It is recommended to create customized Submission Information Package (SIP) templates geared to the specific requirements of the respective RDC.
- As part of such SIP templates, the criteria and guidelines for the submission of contextual materials about the data (e.g. communication such as emails) should be specified further for internal and external data producers. At the same time, guidelines should define which kind of contextual material is part of the Archival Information Package (AIP) and has to be stored in the archiving system.
- The channels through which acquisition takes place have diversified over the years as more teams were established to specifically address certain groups of data producers (e.g. the team International Data Infrastructures, among whose tasks is the acquisition of international surveys). This de-centralized acquisition process, however, makes it necessary to intensify internal communication with the archive staff responsible for the ingest of data. This has partly been accomplished by the establishment of an acquisitions committee. However, in addition, a structured way of tracking planned and ongoing acquisitions should be established.

References

[1] European Framework for Audit and Certification of Digital Repositories (Homepage): http://www.trusteddigitalrepository.eu/

[2] Data Seal of Approval (Homepage): http://www.datasealofapproval.org/en/

[3] CCSDS, Reference Model for an Open Archival Information System (OAIS). Recommended Practice (2012). http://public.ccsds.org/publications/archive/650x0m2.pdf

[4] Schumann, N., and Recker, A., "De-mystifying OAIS compliance: benefits and challenges of mapping the OAIS reference model to the GESIS Data Archive," IASSIST Quarterly, 36, 2 (2013). http://www.iassistdata.org/downloads/iqvol36_2_recker_0.pdf

[5] CCSDS (2012): p. 4-1

[6] GESIS Research Data Centers: http://www.gesis.org/en/institute/competence-centers/

[7] CCSDS, Producer-Archive Interface Methodology Abstract Standard (2004). http://public.ccsds.org/publications/archive/651x0m1.pdf

Author Biographies

Natascha Schumann is affiliated at the Data Archive for the Social Sciences at the GESIS Leibniz Institute for the Social Sciences in Cologne. The main focus of her work is on digital curation of social science research data and audit and certification in this area. Her contact email is natascha.schumann@gesis.org.

Dr. Astrid Recker works at the GESIS Data Archive where she is responsible for the design and delivery of digital preservation workshops for the "Archive and Data Management Training Center." She is also involved in the EU-funded project Data Service Infrastructure for the Social Sciences and Humanities (DASISH). Her contact email is astrid.recker@gesis.org